Research on Image Classification Based on Convolutional Neural Network

Ziling Luo

School of Computing and Data Science, Xiamen University Malaysia, Sepang, Selangor, 43900, Malaysia

JRN2009398@xmu.edu.my

Abstract. The convolutional neural networks (CNNs) are widely used for image classification tasks because CNNs can successfully capture spatial hierarchies and patterns in images. A dataset can be utilized to evaluate the performance of various types of CNNs. To compare the effectiveness of four CNN models for image classification on specific datasets, this study utilizes the MNIST dataset to train four classic CNNs and subsequently compares and evaluates the classification outcomes. The four models are LeNet (LeNet), AlexNet, Visual Geometry Group Network (VGGNet) and Visual Geometry Group Network (ResNet). In order to address the performance of four neural network models in image classification, a controlled experiment is conducted. The results of this study indicate LeNet is the most suitable model on the MNIST dataset. While the other three models also exhibit commendable classification results, they fall short of the overall performance achieved by the LeNet model. The other three models can be used with challenging datasets.

Keywords : Image classification, LeNet-5, AlexNet, Visual Geometry Group Network, Residual Neural Network

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_99

1 Introduction

Image classification is a computer vision class classification task in which feature labels or feature categories are assigned to input images. Whereas an image classification system takes an image and predicts which category it belongs to by using a specific algorithm. Image recognition has numerous applications in object detection and recognition[1], face recognition[2], image search and labeling, medical image recognition, and agricultural image analysis. The image classification is made up of several key components, including feature extraction, image information processing, and judgment or classification[3]. These components work together to enable accurate and efficient image classification. The acquisition of image information comprises getting image data from a variety of sources, such as internet image databases or photographs shot with cameras. The source and quality of the captured images significantly impact the performance of image categorization systems. As a result, ensuring the dependability and high quality of the obtained images is critical to achieve precise categorization results.

After the image processing system completes the acquisition of image data, the subsequent stage of information processing involves image pre-processing and normalization. These two steps encompass a range of procedures, including image cropping, augmentation, color adjustment, and spatial transformations. The most important stage of image classification is feature extraction. In feature extraction, specific algorithms are applied to extract the relevant visual elements, which are one of the important parameters of the image classification system in the image classification work. The deep learning method CNN can be utilized for feature extraction. This is also the method of this study, using different models of CNNs for the classification of images. After extracting the features, the image recognition system proceeds to the judgment or classification processes. Based on the features retrieved in the preceding stage, the system employs a trained classification algorithm or model to assign images to predefined categories or labels.

CNN is a deep learning method[4]. With the use of CNN-based computer vision, it is now possible to carry out some difficult tasks, like facial recognition, self-driving automobiles, self-service checkout lanes, smart medical care, etc. [4].

CNN is inspired by the working principle of visual cells, which have specific response characteristics for perceiving edges and textures in animal brains. This is also the underlying principle of the CNN model in the image recognition process. In

1998, LeCun et al created a convolutional neural network for handwritten postal code classification and proposed the term "convolution," which became the basis for the initial LeNet[5]. Following that, the LeNet-5 model is proposed, which marked an important milestone in the handwritten digit recognition work with the CNN model[5]. In 2012, the AlexNet model made a remarkable breakthrough in the ImageNet image recognition challenge by Alex Krizhevsky et al[6]. The AlexNet model adopts a deep network structure and utilizes the parallel computing capability of Graphics Processing Units (GPU), which greatly speeds up the training process[6]. In 2014, Karen Simonyan and Andrew Zisserman proposed the VGGNet model, a deep CNN model. The VGGNet model employs a small-sized convolution kernel with a deeper network structure, allowing the network to better capture the image's fine properties and obtain good classification results on the ImageNet dataset[7]. VGGNet is also applied to scene recognition. The Google team proposed the GoogLeNet model in 2014, utilizing the Inception module[8]. The GoogLeNet model significantly minimizes the number of parameters by using different convolution kernel sizes and pooling layers in parallel, and it also performs well on the ImageNet dataset[8]. Kaiming He et al. proposed the ResNet model in 2015[9]. The ResNet model addresses the gradient disappearance problem in deep networks by providing residual connections[9].

2 Methods

2.1 LeNet

LeNet is a neural network architecture developed by Yann LeCun. It is a highly efficient convolutional neural network model for recognizing digit handwritten characters. Because of the distinctive architecture of LeNet, it is known as LeNet-5. LeNet-5 has a number of layers. Excluding the input layer, LeNet-5 has a total of 7 layers with some adjustable parameters respectively[5]. LeNet-5 has exceptional learning and recognition capabilities, notably in the field of digit recognition. It is crucial in fostering the development of convolutional neural networks and their application in some areas. As in Fig.1, this structure is the network structure of LeNet-5 which has 5 layers (Combining convolutional layers and pooling layers as one layer). The "5" in the name of LeNet-5 can also be understood as the number of

layers with trainable parameters in the entire network is 5. As the network model becomes deeper, the size of the image decreases as the number of channels increases. There are some differences between the commonly used LeNet-5 structure and the structure proposed by Professor Yann LeCun in his 1998 paper[5], such as the use of activation functions.

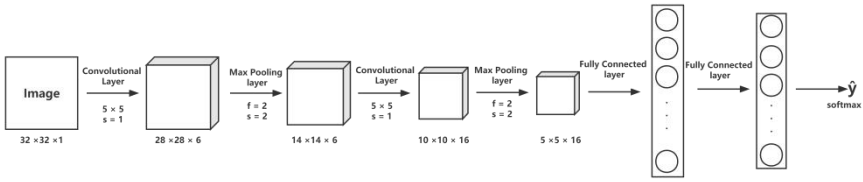


Fig. 1. LeNet-5 model structure (Picture credit: Original)

2.2 AlexNet

Alex Krizhevsky introduces a convolutional neural network architecture named AlexNet in a publication[6]. AlexNet also has many convolutional layers and other network layers like the other models. In addition, to improve the generalization capacity of the network, Local Response Normalization (Local Response Normalization) is used. [6].

AlexNet has some improvements over the neural network before it. Enhancing data augmentation techniques such as random cropping, translation transformation, horizontal flip, color, lighting, and contrast transformation are common data augmentation methods. Dropout effectively prevents overfitting. Replacing the traditional S or T activation function with ReLU is an effective way to improve the model's performance. Overlapping Pooling reduces the overfitting of the system by employing overlapping pooling techniques.

2.3 VGGNet

To enhance the performance of image classification, the depth of the model can be raised. VGGNet has a deep structure with 16-19 convolutional layers, 3 dense layers, and a huge number of parameters[7]. VGGNet employs a relatively small 3x3

convolution kernel that is employed at all network layers[7]. This design option deepens the network, reduces the number of parameters, and enhances model efficiency. This deep structure can collect higher-level characteristics in the image, enhancing recognition accuracy. Simultaneously, VGGNet employs some pooling layers to lower the size of feature map and introduces an activation function (ReLU) to improve the network's nonlinear representation capabilities[7].

2.4 ResNet

ResNet is an image recognition architecture base on deep residual neural networks. ResNet solves the deep network degradation problem through residual learning, allowing deeper networks to be trained, which is a significant development in deep networks. In a classic deep convolutional neural network, an increase in network depth leads to a decrease in accuracy. ResNet addresses this issue by creating residual connections. Each layer's input is directly added to the output in ResNet, establishing a skip connection. In contrast to standard neural networks, the output of each layer must be modified through numerous convolutional layers[9].

3 Results

The Modified National Institute of Standards and Technology database (MNIST) contains a significant number of handwritten digit images. MNIST is frequently used to train image processing systems. This study conducts performance evaluations and analyses on four classic convolutional neural network models using the MNIST dataset.

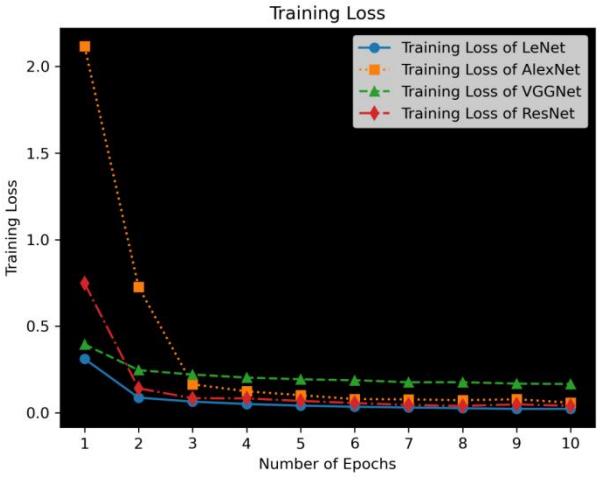
Table 1 includes the hyperparameters which are used in the four CNN models in this study. As can be seen in Table 1, to conduct a controlled experiment, some hyperparameters like the Learning Rate, Epochs, Activation Function, Optimizer, and Loss Function are mostly similar. Only the Kernel Size and Kernel number are

different among these four experiments, which is the main difference between the four models.

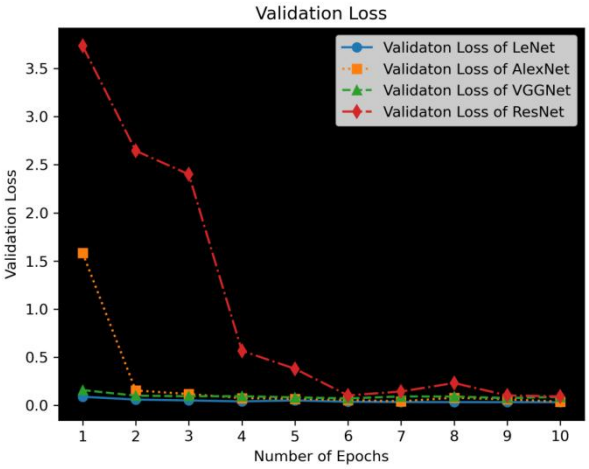
Table 1. A detailed description of the hyperparameter

Hyperparamet	LeNet	AlexNet	VGGNet	ResNet
Learning Rate	0.001	0.001	0.001	0.001
Batch Size	128	128	64	100
Epochs	10	10	10	10
Kernel Size	5 × 5	11 × 11 5 × 5 3 ×	3 × 3	3 × 3
Kernel	22	96 256	1472	1024
Activation	ReLU	ReLU	ReLU	ReLU
Optimizer	Adam	Adam	Adam	Adam
Loss Function	Sparse Categorical Crossentropy	Sparse Categorical Crossentropy	Sparse Categorical Crossentropy	Categorical Crossentropy

Fig. 2 is the comparison of training loss lines and validation loss lines for four different models on MNIST dataset. Table 2 is the loss experimental results of four models on the MNIST dataset. It can be concluded from Fig. 2 and Table 2 that, whether it is training loss or validation loss, LeNet has the lowest loss compared to the other three models. In the ResNet model, the validation loss increases in epoch 8 when other models are basically stable, it has the worst performance in the validation.



(a) Training loss lines of four models on the dataset



(b) Validation loss lines of four models on the dataset

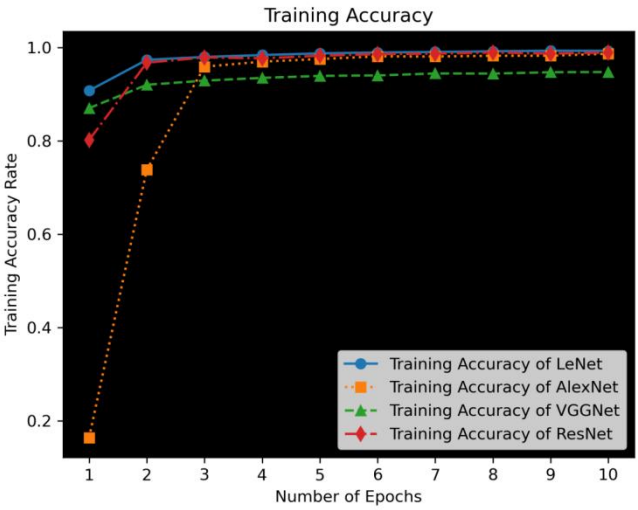
Fig. 2. Comparison of training loss lines and validation loss lines for four different models on the MNIST dataset (Picture credit: Original)

Table 2. Implementation results(loss) of four models on the dataset

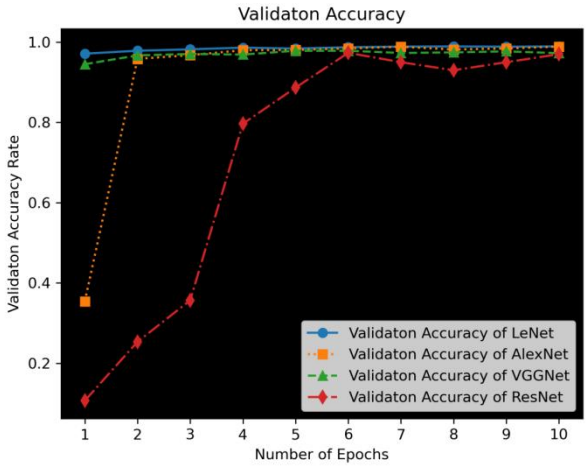
	LeNet	AlexNet	VGGNet	ResNet
Training Loss	0.0220	0.0579	0.1657	0.0392
Validation Loss	0.0338	0.0393	0.0893	0.0936

Fig. 3 is the comparison of training accuracy lines and validation accuracy lines for four different models on the MNIST dataset. Table 3 is the accuracy experimental results of four models on the MNIST dataset. For training accuracy and validation accuracy. LeNet continues to have the best experimental results compared to the other three models, according to the data in Fig. 3 and Table 3.

ResNet shows instability in accuracy lines. In the validation accuracy lines graph, the validation accuracy of the ResNet model decreases in epoch 8 when other models are basically stable. AlexNet and ResNet have changed a lot in the first few epochs.



(a) Training accuracy lines of four models on the dataset



(b) Validation accuracy lines of four models on the dataset

Fig. 3. Comparison of training accuracy lines and validation accuracy lines for four different models on the MNIST dataset (Picture credit: Original)

Table 3. Implementation results(accuracy) of four models on the dataset

	LeNet	AlexNet	VGGNet	ResNet
Training Accuracy	0.9927	0.9865	0.9473	0.9889
Validation Accuracy	0.9893	0.9889	0.9725	0.9700

The images of MNIST are relatively simple and the dataset size is small compared to other more complex image datasets such as ImageNet. AlexNet is a deeper neural network model with more convolutional layers and dense layers than LeNet. Using more complicated models on the comparatively simple MNIST dataset can result in overfitting and training issues, resulting in less effective implementation on MNIST than LeNet.

VGGNet uses a deeper network topology as well as many smaller-sized convolution kernels and pooling layers. The complexity may be too high for simple datasets like MNIST, resulting in overfitting[10].

For ResNet, its design goal is to overcome the deep network training challenge by utilizing residual connections. However, feature extraction does not require an unduly sophisticated model for a simple data set like MNIST, and a simple network structure is sufficient.

Although it is possible to achieve better training results in model training by adjusting the hyperparameters, a more complex model will increase the system load. Except for LeNet, the other three models are not the most suitable for the MNIST dataset.

One issue that occurs when training the AlexNet, VGGNet, and ResNet models on the MNIST dataset is the compatibility between the dataset and the complexity of these models. The architectures of these three models are relatively deep, including numerous convolutional and pooling layers. As a result, the models may encounter the problem of the image size becoming too small after numerous convolutions and pooling procedures when applied to the MNIST dataset, which contains tiny-sized images (28x28 pixels). This issue is not encountered when using the LeNet model because its architecture is specifically designed for the MNIST dataset, which has small-sized images.

4 Conclusion

In this study, four classic convolutional neural network models are applied to the MNIST dataset. By adjusting some hyperparameters, the four models can only achieve high test accuracy and verification accuracy in 10 epochs. LeNet has the best performance when implement on the MNIST compared to the other three CNN models. Since the structures of the other three models are too large, too deep, and too complex, the implementation effect on the MNIST dataset is not as good as the simpler model LeNet. In practice, LeNet can be used on relatively simple data sets, and efficient and accurate results can be obtained. Instead of using large neural network models like AlexNet, VGGNet and ResNet. These three models are applicable to more complex datasets and may be tried to solve specific problems.

References

1. Dhillon, A., Verma, G.K.: Convolutional neural network: a review of models, methodologies and applications to object detection. Prog Artif Intell 9, 85–112 (2020).
2. Almabdy S., Elrefaei L.: Deep Convolutional Neural Network-Based Approaches for Face Recognition. Applied Sciences 9, no. 20: 4397 (2019).
3. Liu N., Wan L., Zhang Y., et al.: Exploiting convolutional neural networks with deeply local description for remote sensing image classification. IEEE Access 6, 11215 – 11228(2018).
4. Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J.: A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. IEEE transactions on neural networks and learning systems, 33(12), 6999–7019(2022).
5. Lecun Y., Bottou L., Bengio Y., Haffner P.: Gradient-based learning applied to document recognition. In: Proceedings of the IEEE, pp. 2278-2324. IEEE(1998).
6. Krizhevsky, A., Sutskever, I., Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks. In: Proceedings of the NIPS, pp. 1097-1105. Curran Associates Inc, Red Hook, NY, USA (2012).
7. Simonyan K., Zisserman A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. 3rd International Conference on Learning Representations (ICLR 2015), Computational and Biological Learning Society, pp. 1–14(2015).
8. Szegedy C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Boston, MA, USA(2015).

9. He, K., Zhang, X., Ren, S., & Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770-778(2015).
10. Zhang X.: The AlexNet, LeNet-5 and VGG NET applied to CIFAR-10. In: 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), pp. 414-419. IEEE, Zhuhai, China(2021).